J Mol Evol (1988) 27:326–335

Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1988

Ancient Origin of Lactalbumin from Lysozyme: Analysis of DNA and Amino Acid Sequences

Ellen M. Prager and Allan C. Wilson

Department of Biochemistry, University of California, Berkeley, California 94720, USA

Summary. Parsimony trees relating DNA sequences coding for lysozymes *c* and α-lactalbumins suggest that the gene duplication that allowed lactalbumin to evolve from lysozyme preceded the divergence of mammals and birds. Comparisons of the amino acid sequences of additional lysozymes and lactalbumins are consistent with this view. When all base positions are considered, the probability that the duplication leading to the lactalbumin gene occurred after the start of mammalian evolution is estimated to be 0.05–0.10. Elimination of the phylogenetic noise generated by fast evolution and compositional bias at third positions of codons reduced this probability to 0.002–0.03. Thus the gene duplication may have long preceded the acquisition of lactalbumin function.

Key words: Parsimony — Evolutionary parsimony — Statistical testing — Phylogenetic noise — Base composition — Transversions — Mammals — Birds — Insects — Mammary gland

Introduction

Lactalbumin is thought, on the basis of amino acid sequence similarities (Shewale et al. 1984) and three-dimensional structure (Stuart et al. 1986), to have arisen from lysozyme *c* by means of a gene duplication. Lysozyme *c* occurs in tissues and secretions (including mammalian milk) of mammals, birds, reptiles, and insects (Jollès and Jollès 1984). Besides being a bacteriolytic enzyme, which appears to function in many creatures as a defense against bacterial infections (Jollès and Jollès 1984), lysozyme has on several occasions been recruited by the gut, where it may have a nutritional role (Stewart et al. 1987 and references therein). Lactalbumin, in contrast, has been found to date only in an organ that is unique to mammals, i.e., the mammary gland, which arose at a known time, approximately 200 million years (Myr) ago. In this organ, lactalbumin binds calcium and interacts with galactosyl transferase so as to promote lactose synthesis (Shewale et al. 1984; Stuart et al. 1986). That a duplication of a lysozyme gene permitted the origin of a protein (lactalbumin) with a new function receives support from observations that genes encoding the two proteins are similar in structure, each comprising four exons and three introns (Qasba and Safaya 1984). Lactalbumin may provide the best example where a new function has developed following a gene duplication.

Nevertheless, uncertainty persists about when the duplication occurred in relation to the time of origin of the new organ. Two models have been proposed to explain when the duplicate genes arose (Fig. 1). Model I proposes an ancient duplication, before the divergence of birds and mammals about 300 Myr ago, i.e., well before the mammary gland arose. This model, supported by phylogenetic analysis of matrices of minimal mutation distances, does not require accelerated evolution along lactalbumin lineages and at the same time can explain why

Offprint requests to: E.M. Prager

lactalbumin is as similar to the conventional[1] lysozymes *c* of birds and reptiles as to those of mammals (Jollès et al. 1977; White et al. 1977). Model II, in contrast, suggests that the lactalbumin gene arose later, i.e., at the start of mammalian evolution, when the mammary gland arose, about 200 Myr ago. Supported by the absence of evidence for either lactose or lactalbumin in nonmammalian vertebrates, model II had to invoke accelerated evolution of lactalbumin relative to lysozyme, which did not appear to be the case (White et al. 1977) despite considerable variation in amounts of change along diverse lactalbumin lineages (Shewale et al. 1984). Shewale et al. (1984) have proposed a modified version of model II, with the duplication along a precursor to the mammalian lineage, i.e., before the mammary gland arose but after the bird–mammal split.

The gene for lysozyme *c* seems to be at least 600 Myr old. This became evident from the amino acid sequences of the 33 amino-terminal residues of lysozyme from three species of moth (Jollès et al. 1979b). The occurrence of lysozyme *c* in invertebrates implies that the last common ancestor of vertebrates and invertebrates, which lived at least 600 Myr ago, possessed the gene for this enzyme. Our phylogenetic analysis (by a distance method) of these insect sequences along with those of vertebrate lysozymes and lactalbumins (Prager and Wilson, unpublished work referred to by Jollès and Jollès 1984) supported model I (Fig. 1) and suggested that the lysozyme–lactalbumin gene duplication might have been as ancient as the vertebrate–invertebrate divergence.

DNA sequences for the entire coding region of three lysozymes *c* (from chicken, human, and moth) and three mammalian lactalbumins[2] along with

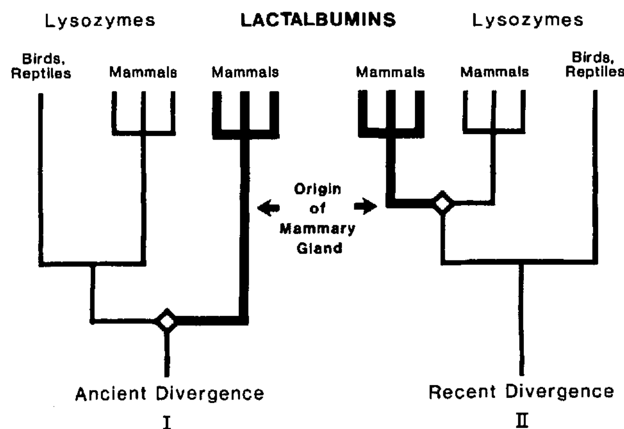

Fig. 1. Alternative models for the origin and divergence of the genes coding for vertebrate lysozyme *c* and α-lactalbumin. Diamonds mark the postulated gene duplication event. Model I proposes that the duplication occurred before the lineage leading to mammals diverged from that leading to birds and reptiles, whereas model II proposes that the duplication occurred relatively late in evolutionary time, along the mammalian lineage. The split between birds and mammals is placed at about 300 Myr ago, the start of mammalian evolution at about 200 Myr ago, and the divergence among different orders of placental mammals (represented by the three mammalian lineages for both proteins in both models) at 60–80 Myr ago (cf. White et al. 1977; Sarich 1985).

amino acid sequences for additional lysozymes and lactalbumins have permitted us to conduct statistical tests that appear to rule out model II.

Materials and Methods

Sequences and Alignments. The six DNA sequences considered, the length (number of amino acids) of the mature protein encoded, and the sources of the sequences are as follows: human lysozyme, 130, Castañón et al. (1988); chicken lysozyme, 129, Jung et al. (1980); moth lysozyme 2, 120, Engström et al. (1985); human and guinea pig lactalbumins, 123, Hall et al. (1982); and rat lactalbumin, 140, Qasba and Safaya (1984). In addition to the DNA encoding the mature proteins, the 12 carboxy-terminal codons of the signal peptides were analyzed; the moth lysozyme sequence was unavailable for more of the signal peptide. The human lysozyme DNA sequence used here was determined from the histiocytic lymphoma cell line U937 and differs by two silent transitions in the signal peptide from the sequence reported for lysozyme DNA from human placenta (Castañón et al. 1988). The chicken lysozyme genomic rather than cDNA sequence was used, because the former agrees entirely with the known amino acid sequence (Jung et al. 1980). A TGA termination codon is used in guinea pig lactalbumin at the same place as in human lactalbumin for the reasons discussed by Hall et al. (1982).

Additional lysozyme sequences considered, at the amino acid level, were from Jollès and Jollès (1984), Stewart et al. (1987), and references therein; additional lactalbumin amino acid sequences appear in Kaminogawa et al. (1984), Shewale et al. (1984), and Beg et al. (1985).

To align vertebrate lysozymes, moth lysozyme, and lactalbumins with one another, it was necessary to introduce three-base, six-base, or nine-base gaps, guided by the information in Hall et al. (1982), Jollès and Jollès (1984), Qasba and Safaya (1984), Shewale et al. (1984), Beg et al. (1985), and Engström et

[1] The term "conventional" applied to vertebrate lysozymes *c* refers to nearly all those whose sequences have been published but provisionally excludes the lysozymes *c* of completely or partially known sequence from horse milk (McKenzie and Shaw 1985), dog milk (Pervaiz and Brew 1986), and pigeon egg white (Rodríguez et al. 1985). These other sequences, termed "unconventional" here, could be the products of gene duplications that preceded the divergence of mammals and birds (Prager, Stewart, and Wilson, unpublished interpretation; see also Rodríguez et al. 1987 and Stewart et al. 1987). The conventional mammalian and bird lysozyme *c* sequences are more similar to one another than are the conventional mammalian sequences to horse and dog milk lysozymes or the conventional bird sequences to pigeon lysozyme. Disregard of these unconventional vertebrate lysozyme *c* sequences (which are available only at the amino acid level) does not interfere with the analysis presented in this report. The genes encoding the conventional bird and mammal lysozymes *c* used for our tree analyses may be regarded as orthologous, implying that their divergence coincided with the divergence of the bird lineage from that leading to mammals

[2] See the Appendix for DNA analyses involving a fourth lactalbumin, from the goat

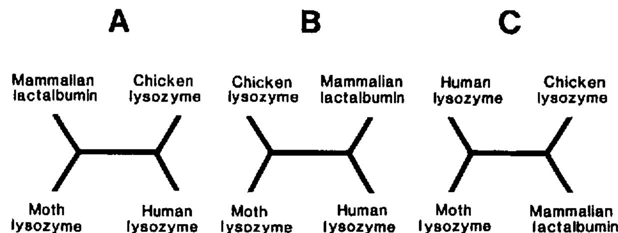

Fig. 2. Three possible unrooted trees relating vertebrate and invertebrate lysozymes *c* to a mammalian α-lactalbumin

al. (1985). To align the fourth exons we used a model not implicating gene rearrangement (cf. Qasba and Safaya 1984). Our alignment of moth lysozyme with the vertebrate lysozymes is as in Engström et al. (1985) except that we put the moth termination codon (TAG) between codons 127 and 128 rather than after codon 129; the rationale is that a single base change with respect to all three lactalbumin sequences (GAG) can produce the moth terminator. This choice has the effects that 9 bp of noncoding DNA in the moth are compared with DNA encoding an amino acid or termination codon in the other sequences and that the overall nucleotide similarity of moth lysozyme DNA to the other DNA sequences rises by an average of 0.8% of the pairwise sequence difference.

Calculations. When counting the number of variable positions or the number of differences between two sequences, we treated additions and deletions as equivalent to nucleotide or amino acid substitutions, as follows: Each addition or deletion of a codon was counted as one amino acid difference; thus, if the gap was three contiguous codons long, it was counted as three amino acid differences or nine nucleotide differences. When doing tree analysis by the parsimony method, we treated each codon gap as one site. Adjacent codon gaps were treated as independent sites. It was further assumed, for character state analyses using amino acid sequences, that a single mutational event sufficed to interchange any amino acid or termination codon and a gap.

Parsimony analysis was done at the DNA level by treating nucleotides as cladistic characters, with all base interchanges considered equally likely. Among four sequences, three bifurcating networks are possible, and Fig. 2 describes the types of four-sequence networks analyzed here. Phylogenetically informative positions (or sites) among four DNA sequences are those where two sequences share one character state and two sequences share another. At any given informative site, therefore, one of the networks in Fig. 2 will require a single mutation to explain the observed diversity among the four sequences and the other networks will each require two events. Swofford's (1985) PAUP program facilitated consideration of the 105 bifurcating networks linking six sequences.

For amino acid sequences, parsimony trees were compared as described before (Jollès et al. 1984). This method uses hypothetical codons and determines the minimum number of mutations required to account for the observed sequence differences. Because sequences were included for which data at the DNA level are unavailable, the analyses of amino acid sequences were done as if no DNA data were known for any of the four proteins included in a particular comparison. Analogous to the treatment of DNA results, a tree that requires fewer base substitutions than another tree at a given position wins at that position.

With the one-tailed binomial test we calculated the probability that two trees, X and Y, are equivalent. For each pair of trees or networks, this test counts the number of positions at which tree X wins, i.e., explains the observed sequence diversity with fewer mutational events, and the number of positions at which tree Y wins. The probability *P* of accepting the null hypothesis, that tree X is no better than tree Y and that the observed score does not differ from random expectation, is

$$P = (½)^n \sum_{X=r}^{n} \binom{n}{X}, \qquad (1)$$

where n is the number of positions compared, r is the number favoring tree X over tree Y (with $r \geq n/2$), and

$$\binom{n}{X} = \frac{n!}{X!(n - X)!}.$$

DNA sequences were compared also with the evolutionary-parsimony method (Lake 1987). This method divides the sites used to assess branching order into two categories, parsimony and background. Parsimony sites favor one of the three networks linking four sequences and background sites are subtracted as a correction factor. The statistical significance of evolutionary-parsimony trees was evaluated with the chi-square test in Eq. 9 of Lake (1987).

To quantify differences in base composition between two sequences, we used a difference index, D, defined in Eq. 2,

$$D = |G_1 - G_2| + |A_1 - A_2| + |T_1 - T_2| + |C_1 - C_2|, \qquad (2)$$

where G_1, A_1, T_1, and C_1 are the compositions in percent of the four bases in sequence 1 and G_2, A_2, T_2, and C_2 are the corresponding quantities for sequence 2. Compositions and D values were computed without correction for the slight variation in sequence length (from 408 to 429 bp) and the occurrence of gaps among the six DNA sequences.

Results

Sequence Comparisons

The aligned base sequences for the coding regions of three lysozymes (human, chicken, and moth) and three lactalbumins (human, guinea pig, and rat) are shown in Fig. 3. The numbers of DNA and amino acid differences observed at the 432 coding positions compared for all possible pairs of the six sequences are given in Table 1. Consistent with previous observations (Fig. 1), human and chicken lysozymes are more similar to each other than either is to the mammalian lactalbumins, and, moreover, these two lysozymes are equally different from the lactalbumins. The moth lysozyme sequence is about equally different from the other five sequences (particularly at the DNA level); furthermore, the magnitude of this difference is greater than that between vertebrate lysozymes and lactalbumins. In addition, the three lactalbumin sequences in the table are most similar to one another.

For all 15 pairwise comparisons in Table 1, the differences are greatest at the third position of codons and least at the second position. Indeed, save among the lactalbumins, differences at the third position, where silent changes predominate, approach saturation: At third positions the average point-mutational difference among the three lactalbumins, which diverged 60–80 Myr ago, is 33%. By 300 Myr,

```
                                  -10             -5              -1  1           5               10
Human Lysozyme        CTG GGG CTT GTC CTC CTT TCT GTT ACG GTT CAA GGC AAG GTC TTT GAA AGG TGT GAG TTG GCC AGA ACT
Chicken Lysozyme      T.. .T. ... TG. T.. ..G C.C C.G G.T .C. .TG ..G ..A ... ... .G. C.A ... ... C.. ..A GCG G..
Moth Lysozyme         TGC C.T TCG TGG .AG T.C G.. T.G CAT TGC G.T .CG ..A CGT ..C ACG ..A ..C .G. ..A .TG CAG GAG
Human Lactalbumin     T.C CT. G.G .G. A.. ..G .TC CC. G.C A.C .TG .C. ... CAA ..C AC. .AA ... ... C.. T.. CAG CTG
GP Lactalbumin        T.. CT. G.G .G. A.. ..G .T. CC. G.C ..G ..G .C. ... CAA C.. ACC .AA ... .C. C.. T.T CAT GAG
Rat Lactalbumin       T.C CTA GCG TGT A.T TCG CTG CC. G.C T.. ... .C. .CA .AG ... AC. .AA ... ... G.. T.. CAC G.C

                         15              20              25              30              35
Human Lzm     CTG AAA AGA TTG GGA ATG GAT GGC TAC AGG GGA ATC AGC CTA GCA AAC TGG ATG TGT TTG GCC AAA TGG GAG AGT
Chicken Lzm   A.. ..G C.T CAC ... C.T ... AA. ..T C.. ... TA. ... ..G .G. ... ... G.. ... GCC ..A ... .TC ... ...
Moth Lzm      ..T .GG ... CGA ..C T.C ... .AA ACT TT. --- --- --- A.G AGT ... ... G.C ..C C.T .T. G.G AAC ..A ..C
Human Lac     ... ... --- --- .AC ..A ... ..T ..T G.A ..C ... GCT T.G C.T G.A .T. ..C ... ACC ATG TTT CAC ACC ...
GP Lac        T.. ..C --- --- .AC C.T .CC ... ... C.A .AC ... .CT T.G C.T G.A ... C.C ... A.C ATA TTT CAT ATC ...
Rat Lac       A.T G.. --- --- .AC ... ... ... ..T CAA ..C ... ... T.G CTT G.A ... .CC ... G.T TTA TTC CAC ACC ...

                         40              45              50              55              60
Human Lzm     GGT TAC AAC ACA CGA GCT ACA AAC TAC AAT GCT GGA GAC AGA AGC ACT GAT TAT GGG ATA TTT CAG ATC AAT AGC
Chicken Lzm   AAC .T. ... ..C .AG ... ... ... CGT ..C A.C --- ..T G.G ..T ..C ..C ..C ..A ..C C.A ... ... ..C ...
Moth Lzm      ..A CGG TTT ..C GAT AAA .TC GGT A.A GT. AAC AAG A.. G.. TCT CGA ..C ..C ..C C.C ..C ... ... ... GA.
Human Lac     ... ..T G.. ... .A. ..C .T. GTT G.A ..C --- --- A.T GA. ... ..G ..A ... ..A C.C ..C ... ... .G. .AT
GP Lac        ... ..T G.. ... .A. ..C .T. GTG A.A ... --- --- AGT GAC CA. .AA ..G ..C ..A C.T ..C ... ..A ... GAT
Rat Lac       ..C ..T G.. T.. .A. ... .TC GT. A.G ..C --- --- A.T G.C ... ..A ..G ... ..A C.C ..C ... ... .G. .A.

                         65              70              75              80              85
Human Lzm     CGC TAC TGG TGT AAT GAT GGC AAA ACC CCA GGA GCA GTT AAT GCC TGT CAT TTA TCC TGC AGT GCT TTG CTG CAA
Chicken Lzm   ... .GG ... ..C ..C ... ... .GG ... ... ..C T.C AGG ..C CTG ..C A.C A.C C.G ... TCA ..C C.. ... AGC
Moth Lzm      AAA ... ... ..C .G. A.G ..A TCC ..T ..T ... --- AAG G.. --- ..C A.C G.G A.T ..T .A. CAG C.A ... ACT
Human Lac     AAG CTT ... ..C ..G AGC A.. C.G GT. ..T CAG T.. AGG ..C AT. ... G.C A.C ... ..T GAC AAG ..C ... G.T
GP Lac        AAA G.T .TC ... G.G AGC A.. .CG ..T GTT CA. T.. AGG ..C ATT ... G.C A.T ... ... GAC AAG C.C ... G.T
Rat Lac       A.A A.T ... ..C ..G AG. A.. G.G TT. ..C .AG T.. .AG ..C AT. ... G.C A.C ... ..T GAC AAG ..C ..T G.T

                         90              95              100             105             110
Human Lzm     GAT AAC ATC GCT GAT GCT GTA GCT TGT GCA AAG AGG GTT GTC CGT GAT CCA CAA GGC ATT AGA GCA TGG GTG GCA
Chicken Lzm   TCA G.. ..A A.A .CG AGC ..G AAC ..C ..G ... .A. A.C ... A.C ... GG. A.C ... ..G .AC ..G ... ..C ..C
Moth Lzm      ..C G.. ..T AGC .TG ..A .CT A.G ..C ..G ... .A. A.. TA. AAA --- .GC ..C AAG T.. GAC ..T ... TAC .G.
Human Lac     ... G.. ..T A.. ... .AC A.. ATG ... ..C ... .A. A.C C.G --- ... ATT A.. ..A ... GAC TAC ... T.. ..C
GP Lac        ... G.. C.T A.. ... .AC A.. ATG ... .TC ... .A. A.C C.G --- ... ATC A.. ..A ... GAC TAC ... T.. ..C
Rat Lac       ... G.G C.T ..G ... .AC A.. .TA ... ..C ... .A. A.C ..G --- .C. ATC A.. ..G ..C GAC TAC ... AA. ..T

                         115             120             125             129
Human Lzm     TGG AGA AAT CGT TGT CAA AAC AGA GAT GTC CGT CAG TAT GTT CAA GGT TGT --- GGA GTG TAA
Chicken Lzm   ... C.C ..C ..C ..C A.G GG. .CC ..C ... .AG GC. .GG A.C AG. ..C ..C ... C.G C.. .G.
Moth Lzm      ... .A. ... .AC ... ... C.T --- --- .GA .TG .CA G.. A.. AGC .AC ... TAG A.. CGA CTT
Human Lac     CAT .A. GCC .TC ..C ACT G.G .AG --- C.G GAA ... .GG C.. --- --- ... GAG AAG T.. .G.
GP Lac        CAC .A. CCA .TG ..C TCT G.. .AG --- C.G GAG ... .GG TAC --- --- ..C GAG .C. CA. .G.
Rat Lac       CAC .AG CCC ATG ..C TCT G.G .AG --- C.G GAA ... .GG CGC --- --- ... GAG AAG CC. GG.
```

Fig. 3. DNA sequences for human, chicken, and moth lysozymes *c* and three mammalian α-lactalbumins. The bases in the human lysozyme sequence are arranged in coding triplets. Bases in other sequences appear only when different from that in human lysozyme; sequence identity is indicated by a dot. Dashes indicate deletions in one or more sequences relative to other sequences shown. The termination codon in each sequence is underlined; rat lactalbumin, due to a single base change abolishing the TGA stop codon shown for the other lactalbumins, has 17 more amino acids than the human and guinea pig proteins (cf. also Beg et al. 1985). Codons are numbered 1–129 corresponding to the amino acids of mature chicken lysozyme, following a previous convention (Jollès et al. 1984); human lysozyme has one amino acid inserted between codons 47 and 48. Codons −1 to −12 correspond to the 12 carboxy-terminal residues of the signal peptide. [Guinea pig lactalbumin codon 65 is GAG (R.K. Craig, personal communication); the GAC at this position in Hall et al. (1982) represents a typographical error.] Of the 432 bp within the region shown, 359 are variable; asterisks mark the 191 sites that are phylogenetically informative when all six sequences are considered simultaneously in a parsimony analysis. GP, guinea pig.

when chicken and human lysozymes diverged, the difference has reached 57%, and it increases only slightly more, to 69%, when moth and vertebrate lysozymes, which diverged over 600 Myr ago, are compared. Such saturation would be expected if the average rate of divergence at synonymous sites observed for nuclear genes (about 1% per Myr; Wilson et al. 1987) applies to lysozyme.

For eight of the nine comparisons of lactalbumins with lysozymes, transversions outnumber transitions (Table 1). This is true also for the comparison of moth lysozyme with vertebrate lysozymes. By contrast, the differences between more similar sequences are predominantly transitions. These observations fit with the view that transitions occur more often than transversions and that the more dissimilar two sequences are, the more likely it is that transversions have erased the record of transitions (DeSalle et al. 1987).

Differences in base composition among the six sequences are shown in Fig. 4, and the compositional differences are quantified in Table 2. Chicken ly-

Table 1. Numbers of differences among the lysozyme and lactalbumin sequences in Fig. 3

Sequences compared	Total bp	Transitions	Transversions	Gaps: Total	Gaps: Contiguous	Codon position 1	Codon position 2	Codon position 3	Amino acids
Human lzm vs chicken lzm	170	88	79	3	1	50	36	81	60
Human lzm vs human lac	220	102	91	27	6	64	50	79	93
Human lzm vs guinea pig lac	232	94	111	27	6	69	56	80	96
Human lzm vs rat lac	228	96	105	27	6	66	52	83	95
Chicken lzm vs human lac	218	86	108	24	6	65	51	78	91
Chicken lzm vs guinea pig lac	231	81	126	24	6	63	61	83	97
Chicken lzm vs rat lac	226	84	118	24	6	65	58	79	97
Moth lzm vs human lzm	250	102	121	27	6	67	60	96	95
Moth lzm vs chicken lzm	243	87	126	30	7	63	61	89	95
Moth lzm vs human lac	252	94	116	42	8	64	63	83	101
Moth lzm vs guinea pig lac	257	91	124	42	8	64	61	90	103
Moth lzm vs rat lac	258	88	128	42	8	67	60	89	106
Human lac vs guinea pig lac	83	47	36	0	0	23	21	39	40
Human lac vs rat lac	91	53	38	0	0	30	19	42	43
Guinea pig lac vs rat lac	115	69	46	0	0	32	29	54	54

All columns except the rightmost refer to DNA sequence differences. Under gaps, the first column gives the total number of nucleotide differences when each deleted codon is counted as 3 bp differences, and the second column gives the number of gaps when adjacent deleted nucleotides (3–9 bp) are counted as a single gap. In the comparisons of lactalbumin with moth lysozyme, the deletions in the former at codon 100 and the latter at 101 in Fig. 3 have been counted as one deletion for computing the entry in the contiguous column. Differences at individual codon positions refer to point-mutational differences only. *Abbreviations:* lzm = lysozyme *c*; lac = α-lactalbumin

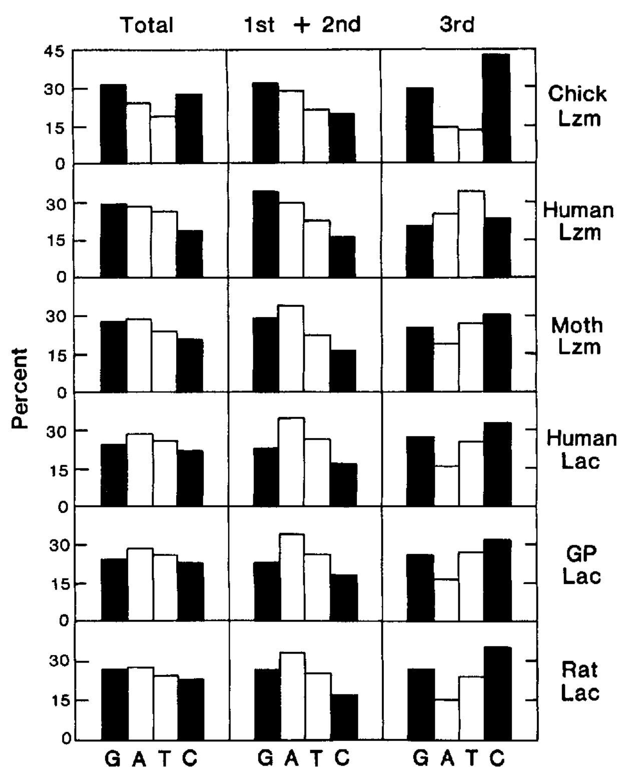

Fig. 4. Base compositions in percent of animal lysozymes *c* and α-lactalbumins. The six histograms on the left give compositions for the complete sequences in Fig. 3; compositions were computed for 429 bp of human lysozyme, 426 bp of chicken lysozyme, and 408 bp of the remaining four sequences. The middle and right histograms, respectively, provide percent compositions at the first plus second and at the third positions of the triplets in Fig. 3. Chick, chicken; GP, guinea pig.

sozyme stands out most from the other sequences (Fig. 4). Moth lysozyme is nearly as similar in overall base composition to the three lactalbumins as these are to one another; the average difference index (D) is 7 compared to 4 among the lactalbumins (Table 2). The biggest differences in base composition occur at third positions of codons, where D is on average twice as great as at first and second positions (Table 2). Human and chicken lysozymes are most different from each other (D = 63) and also more different from the other four sequences

Table 2. Differences in base compositions among animal lysozymes *c* and α-lactalbumins

Sequences compared	Human lzm	Chicken lzm	Human lac	Guinea pig lac	Rat lac	Moth lzm
Human lzm	—	23/6	9/21	10/21	9/14	7/9
Chicken lzm	63	—	24/23	24/21	19/17	20/12
Human lac	35	27	—	1/2	5/7	7/12
Guinea pig lac	31	32	5	—	5/7	8/12
Rat lac	40	23	6	9	—	5/5
Moth lzm	26	36	9	4	13	—

Difference index (D) was computed with Eq. 2 from the base compositions in percent in Fig. 4. D values for all positions considered together, followed by D for the first plus second codon positions, are shown above the diagonal; D values for third codon positions appear below the diagonal. (The computations were done using an additional significant figure for base compositions and the D values then rounded to the nearest whole number.) *Abbreviations:* As in Table 1

(mean D = 31) than those sequences are from one another (mean D = 8). The marked preference for G and C at third positions in chicken lysozyme (73%) contrasts with the preference for T and A at these positions (58%) in human lysozyme (Fig. 4).

Tree Analysis and Statistical Testing

Parsimony

Among the 432 positions shown in Fig. 3, 73 sites are invariant in the six sequences and 359 are variable. Of the latter, 191 are phylogenetically informative for a parsimony analysis. If fewer than six sequences are considered simultaneously, a subset of the 191 sites is phylogenetically informative for parsimony analysis.

The results of using four DNA sequences at a time to compare trees A–C in Fig. 2 are summarized in Table 3. When all positions and all mutations are considered (first set of comparisons in the table), tree A, which associates chicken and human lysozymes, is in each case better than tree B or C. For example, when the lactalbumin used is human, tree A wins at 30 sites, whereas tree B, which associates mammalian lactalbumins with mammalian lysozymes, wins at only 18 sites. Random expectation is that each tree would win at 24 sites. The probability that the observed result would occur by chance is 0.056. Tree A is superior at the 5% level of significance for five of the six comparisons at the top of Table 3. Trees B and C, in contrast, are always statistically equivalent to each other, which means that it is no better to ally mammalian lactalbumin with mammalian lysozyme (tree B) than with bird lysozyme (tree C). Similarly, when all mutations at all positions except those involving gaps are considered, tree A is favored over trees B and C (P = 0.014–0.092).[3]

The remainder of Table 3 reveals that more statistically significant information emerges when attention is confined to positions at which change occurs slowly. That is, when only first and second positions of codons are considered, tree A is always favored over trees B and C at the 5% level of significance. Indeed, there is no discrimination among the three trees when only third positions are examined (P = 0.33–0.59). Similarly, positions differing by transitions overall provide less support for tree A (P = 0.18–0.50 at all positions) than do those differing by transversions (P = 0.01–0.20).

Results from comparisons of amino acid sequences, though based on fewer informative sites, likewise support tree A (Table 4), but not as con-

[3] The results of comparing all six DNA sequences at once likewise show that trees congruent with the branching order of tree A in Fig. 2 are more parsimonious than those congruent with trees B and C; however, the P values were slightly higher than those obtained for four sequences (see the Appendix for details)

Table 3. Statistical comparison of parsimony trees for three animal lysozymes and one lactalbumin according to various categories of sites[a]

Position, mutation, and lactalbumin	Sites favored by tree			Probability[b]	
	A	B	C	A vs B	A vs C
All positions, all mutations					
Human	30	18	17	0.056	0.039*
Guinea pig	30	16	13	0.027*	0.0069*
Rat	31	18	14	0.043*	0.0080*
All positions except gaps					
All mutations					
Human	28	18	17	0.092	0.068
Guinea pig	28	16	13	0.048*	0.014*
Rat	29	18	14	0.072	0.016*
Transversions					
Human	14	9	6	0.202	0.058
Guinea pig	16	9	5	0.115	0.013*
Rat	17	7	5	0.025*	0.0085*
Transitions					
Human	14	9	11	0.202	0.345
Guinea pig	12	7	8	0.180	0.252
Rat	12	11	9	0.500	0.332
First + second positions					
All mutations					
Human	17	6	7	0.017*	0.025*
Guinea pig	19	4	4	0.0013*	0.0013*
Rat	18	7	4	0.022*	0.0022*
Transversions					
Human	7	3	4	0.172	0.274
Guinea pig	8	3	2	0.113	0.055
Rat	7	1	3	0.035*	0.172
Transitions					
Human	10	3	3	0.046*	0.046*
Guinea pig	11	1	2	0.0032*	0.011*
Rat	11	6	1	0.166	0.0032*
Third position					
All mutations					
Human	11	12	10	0.500	0.500
Guinea pig	9	12	9	0.332	0.593
Rat	11	11	10	0.584	0.500
Transversions					
Human	7	6	2	0.500	0.090
Guinea pig	8	6	3	0.395	0.113
Rat	10	6	2	0.227	0.019*
Transitions					
Human	4	6	8	0.377	0.194
Guinea pig	1	6	6	0.063	0.063
Rat	1	5	8	0.109	0.020*

[a] The three trees compared (A, B, and C) appear in Fig. 2. The indicated lactalbumin was tested with human, chicken, and moth lysozymes. For the first set of comparisons (all positions, all mutations), all 432 positions in Fig. 3 were considered. For the remainder of the table the sites involving gaps were omitted

[b] The probability computed using Eq. 1 that trees A and B or A and C are equivalent, i.e., statistically equally parsimonious in accounting for the observed sequence diversity. P values <0.05 are marked with asterisks (*)

Table 4. Statistical evaluation of parsimony scores for evolutionary trees relating amino acid sequences of three animal lysozymes *c* and one α-lactalbumin[a]

Bird lysozyme[b]	Mammal lysozyme	Lactalbumin	Informative sites (no.)	Trees compared: A vs B Score	A vs B P	A vs C Score	A vs C P	B vs C Score	B vs C P
1) Chicken	Human	Human	26	14:7	0.095	14:6	0.058	6:5	0.500
2) Chicken	Human	Human	31	17:8	0.054	16:7	0.047*	7:7	0.605
3) Chicken	Cow 2[c]	Cow	28	16:8	0.076	16:4	0.0059*	8:4	0.194
4) Chicken	Cow 2	Rat	30	16:8	0.076	16:7	0.047*	7:6	0.500
5) Chachalaca	Baboon	Camel	26	17:6	0.017*	17:4	0.0036*	5:3	0.363
6) Duck 1	Langur	Horse	29	17:3	0.0013*	17:9	0.084	3:9	0.073
7) Guinea fowl	Human	Kangaroo	29	15:10	0.212	15:5	0.021*	9:4	0.133
8) Chicken	Rat	Rat	29	13:10	0.339	13:7	0.132	9:6	0.304
9) Chicken	Rat	Human	28	12:10	0.416	12:7	0.180	9:6	0.304

[a] In each example moth lysozyme was tested with the three tabulated vertebrate proteins and the three trees in Fig. 2 compared (with the bird and mammal lysozymes, respectively, placed where chicken and human lysozymes are in the figure). Only the mature proteins were considered except in example 2, where 12 amino acids of the signal peptides (Fig. 3) were also included. For each pair of trees, the score with the number of sites at which each tree wins and *P* computed using Eq. 1 are shown; asterisks mark *P* values <0.05. At a few amino acid positions (two sites in example 2, one in examples 1, 4, 5, and 7–9), two of the three trees are equally parsimonious and superior to the third tree; therefore the numbers of sites individually favored by trees A, B, and C appear to add up to more than the indicated number of informative sites

[b] All bird lysozyme sequences in the table were assumed to have asparagine at residue 103, for the reasons given by Jollès et al. (1979a)

[c] The cow 2 stomach lysozyme sequence has been revised (J. Jollès and P. Jollès, personal communication): histidine, not lysine, occurs at position 97

vincingly as do the DNA sequences. In seven of the 18 tests, tree A wins over trees B or C at the 5% level of significance, whereas trees B and C are never superior statistically to tree A.

Evolutionary Parsimony

The evolutionary-parsimony method was also applied to sets of four DNA sequences to try to choose among trees A, B, and C (Table 5). In marked contrast to the results in Tables 3 and 4, this approach (1) never discriminates among the three trees when the more slowly evolving first and second positions are considered, (2) sometimes distinguishes among the trees when all positions or third positions alone are analyzed, and (3) does not consistently favor the same tree, with B winning when human or guinea pig lactalbumin is included in the comparison and A winning when rat lactalbumin is used.

Table 5. Statistical comparison of evolutionary-parsimony trees for three animal lysozymes and one lactalbumin

Lactalbumin	Tree	Ratio of parsimony to background scores at various codon positions: All	First + second	Third
Human	A	17:12	8:5	9:7
Human	B	11:4†	4:3	7:1*
Human	C	7:5	5:2	2:3
Guinea pig	A	17:15	8:9	9:6
Guinea pig	B	10:2‡	3:2	7:0**
Guinea pig	C	7:3	3:2	4:1
Rat	A	20:8‡	7:4	13:4*
Rat	B	7:6	1:4	6:2
Rat	C	6:6	3:3	3:3

Trees A, B, and C are those in Fig. 2; as in Table 3, the indicated lactalbumin was tested with human, chicken, and moth lysozymes. *P* values in the chi-square test with one degree of freedom are marked as follows: †, <0.10; *, <0.05; ‡, <0.025; **, <0.01. When all data are considered simultaneously using the formulas applicable to the situation where each comparison does not represent an independent measurement (Lake 1987), the A and B trees, respectively, had *P* values of <0.10 and <0.025 for all positions and also for third positions alone

Discussion

Strengths and Weaknesses of Parsimony and Evolutionary Parsimony

When attention was confined to the slowly changing first and second positions of codons (Table 3), the standard parsimony test consistently indicated that tree A (Fig. 2) is superior to trees B and C for describing the relationship of lactalbumin to animal lysozymes. It was to be expected that, because of the phylogenetic noise created by saturation effects, the superiority of tree A would not be evident when the faster-changing third positions were considered. However, it was not expected that tree C would be better than tree A for third positions differing by transitions (Table 3). To help explain this apparent failure of the parsimony method, we draw attention to the marked differences in base composition at third positions (Table 2 and Fig. 4). It would appear that chicken and human lysozyme sequences have

been subjected to strong selection for different base compositions at such sites (cf. Bernardi and Bernardi 1986). As a consequence, the two sequences would be expected to differ often at such sites even if A were the true tree. The importance of confining attention when building a tree to sites at which there are not enormous differences in base composition (cf. Preparata and Saccone 1987) is underscored by our study.

Evolutionary parsimony was designed for situations where parsimony might fail and requires long sequences to function optimally. Because parsimony did select one tree and the sequences here are short, it is premature to regard the disappointing performance of evolutionary parsimony in our case as indicating that this method is generally inferior to the standard parsimony method.

Ancient Origin of Lactalbumin from Lysozyme

Parsimony analyses of DNA and amino acid sequences for vertebrate and insect lysozymes *c* and for mammalian lactalbumins have provided statistically significant evidence in support of an ancient gene duplication, occurring before the bird–mammal divergence, to produce the genes now encoding mammalian lactalbumins. Although such a demonstration makes this model plausible, we should not consider it established until lactalbumin-like genes (or pseudogenes) or proteins are shown to be present in nonmammals (Wilson et al. 1977).[4]

If lactalbumin and lactose prove to be detectable only in mammalian milk, a long period probably elapsed between the time of the gene duplication and the acquisition of lactalbumin's new function (in lactose synthesis). During this period, the product of the duplicate gene could have had a role in calcium metabolism (Stuart et al. 1986) or in modifying the specificity of glycosyl transferases in ways that affect the pattern of glycosylation on cell surfaces (Spicer et al. 1987). A crystallographic study (Stuart et al. 1986) of the residues in lactalbumin critical for binding calcium (aspartic acid at positions 85, 90, and 91 in Fig. 3) encouraged the view that there was an intermediate stage at which the protein could perform bacteriolysis and bind calcium. Nitta et al. (1987) have since reported calcium binding by horse milk lysozyme, which has the three critical aspartyl residues. The detection (McKenzie and White 1987) of traces of bacteriolytic activity associated with lactalbumin further supports the idea of both functions in one molecule. [Hayssen and Blackburn (1985) imagine a physiologically intermediate stage in the evolution of lactation at which maternal skin glands secreted a fluid "protomilk," which, rather than being nutritive, protected the newborn animal from bacteria.]

Among lysozymes *c* whose sequences have been published, only the unconventional horse milk and pigeon egg white lysozymes[1] resemble lactalbumin in having aspartyl residues at all three positions critical for calcium binding. This observation prompts the question whether lactalbumin may be more related to calcium-binding (unconventional) than to conventional vertebrate lysozymes *c*. However, lactalbumins are not more similar in their amino acid sequences to the unconventional lysozymes. Definitive tests of this possibility will become possible when more sequences are available for calcium-binding lysozymes *c*.

[4] One reason for making this point is that, in theory, the parsimony method could be misleading if the case of molecular evolution under study were dominated by positive Darwinian selection rather than by the drift of neutral mutations. Thus, in the situation just described it is possible that at an early stage in mammalian evolution selection for the new function could have erased the record of lactalbumin's common ancestry with mammalian lysozymes implied in model II (Fig. 1). At the same time, the parallel acquisition of warm-bloodedness in birds and mammals may have produced convergence between bird and mammal lysozymes as these animals faced bacteria exploiting the warm-blooded niche. Although functional convergence is usually unaccompanied by sequence convergence, the possibility that sequence convergence can occur in such a circumstance has been raised (Stewart et al. 1987)

Acknowledgments. We thank M.J. Castañón for providing human lysozyme cDNA sequences before publication, M. Stoneking for computer assistance, C.-B. Stewart and J.A. Lake for helpful discussion, and the National Institutes of Health and National Science Foundation for support.

References

Beg OU, von Bahr-Lindström H, Zaidi ZH, Jörnvall H (1985) The primary structure of α-lactalbumin from camel milk. Eur J Biochem 147:233–239

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Castañón MJ, Spevak W, Adolf GR, Chlebowicz-Sledziewska E, Sledziewski A (1988) Cloning of human lysozyme and expression in the yeast *Saccharomyces cerevisiae.* Gene 66: 223–234.

DeSalle R, Freedman T, Prager EM, Wilson AC (1987) Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila.* J Mol Evol 26:157–164

Engström Å, Xanthopoulos KG, Boman HG, Bennich H (1985) Amino acid and cDNA sequences of lysozyme from *Hyalophora cecropia.* EMBO J 4:2119–2122

Hall L, Craig RK, Edbrooke MR, Campbell PN (1982) Comparison of the nucleotide sequence of cloned human and guinea-pig pre-α-lactalbumin cDNA with that of chicken pre-lysozyme cDNA suggests evolution from a common ancestral gene. Nucleic Acids Res 10:3503–3514

Hayssen V, Blackburn DG (1985) α-lactalbumin and the origins of lactation. Evolution 39:1147–1149

Jollès P, Jollès J (1984) What's new in lysozyme research? Mol Cell Biochem 63:165–189

Jollès J, Sen A, Prager EM, Jollès P (1977) Structural data concerning reptilian (tortoise) egg lysozyme. J Mol Evol 10: 261–264

Jollès J, Ibrahimi IM, Prager EM, Schoentgen F, Jollès P, Wilson AC (1979a) Amino acid sequence of pheasant lysozyme. Evolutionary change affecting processing of prelysozyme. Biochemistry 18:2744–2752
Jollès J, Schoentgen F, Croizier G, Croizier L, Jollès P (1979b) Insect lysozymes from three species of Lepidoptera: their structural relatedness to the c (chicken) type lysozyme. J Mol Evol 14:267–271
Jollès P, Schoentgen F, Jollès J, Dobson DE, Prager EM, Wilson AC (1984) Stomach lysozymes of ruminants. II. Amino acid sequence of cow lysozyme 2 and immunological comparisons with other lysozymes. J Biol Chem 259:11617–11625
Jung A, Sippel AE, Grez M, Schütz G (1980) Exons encode functional and structural units of chicken lysozyme. Proc Natl Acad Sci USA 77:5759–5763
Kaminogawa S, McKenzie HA, Shaw DC (1984) The amino acid sequence of equine α-lactalbumin. Biochem Int 9:539–546
Kumagai I, Tamaki E, Kakinuma S, Miura K-I (1987) Molecular cloning and sequencing of cDNA encoding goat pre α-lactalbumin. J Biochem 101:511–517
Lake JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191
McKenzie HA, Shaw DC (1985) The amino acid sequence of equine milk lysozyme. Biochem Int 10:23–31
McKenzie HA, White FH Jr (1987) Studies of a trace cell lytic activity associated with α-lactalbumin. Biochem Int 14:347–356
Nitta K, Tsuge H, Sugai S, Shimazaki K (1987) The calcium-binding property of equine lysozyme. FEBS Lett 223:405–408
Pervaiz S, Brew K (1986) Purification and characterization of the major whey proteins from the milks of the bottlenosed dolphin (*Tursiops truncatus*), the Florida manatee (*Trichecus manatus latirostris*), and the beagle (*Canis familiaris*). Arch Biochem Biophys 246:846–854
Preparata G, Saccone C (1987) A simple quantitative model of the molecular clock. J Mol Evol 26:7–15
Qasba PK, Safaya SK (1984) Similarity of the nucleotide sequences of rat α-lactalbumin and chicken lysozyme genes. Nature 308:377–380
Rodríguez R, Menéndez-Arias L, González de Buitrago G, Gavilanes JG (1985) Amino acid sequence of pigeon egg-white lysozyme. Biochem Int 11:841–843
Rodríguez R, Menéndez-Arias L, González de Buitrago G, Gavilanes JG (1987) Structure of the pigeon lysozyme and its relationship with other type *c* lysozymes. Comp Biochem Physiol 88B:791–796
Sarich VM (1985) Rodent macromolecular systematics. In: Luckett WP, Hartenberger J-L (eds) Evolutionary relationships among rodents. A multidisciplinary analysis. Plenum, New York, pp 423–452
Shewale JG, Sinha SK, Brew K (1984) Evolution of α-lactalbumins. The complete amino acid sequence of the α-lactalbumin from a marsupial (*Macropus rufogriseus*) and corrections to regions of sequence in bovine and goat α-lactalbumins. J Biol Chem 259:4947–4956
Spicer SS, Erlandsen SL, Wilson AC, Hammer MF, Hennigar RA, Schulte BA (1987) Genetic differences in the histochemically defined structure of oligosaccharides in mice. J Histochem Cytochem 35:1231–1244
Stewart C-B, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401–404
Stuart DI, Acharya KR, Walker NPC, Smith SG, Lewis M, Phillips DC (1986) α-lactalbumin possesses a novel calcium binding loop. Nature 324:84–87
Swofford DL (1985) Phylogenetic analysis using parsimony (PAUP), Version 2.4. Illinois Natural History Survey, Champaign, IL
White TJ, Mross GA, Osserman EF, Wilson AC (1977) Primary structure of rat lysozyme. Biochemistry 16:1430–1436
Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46:573–639
Wilson AC, Ochman H, Prager EM (1987) Molecular time scale for evolution. Trends Genet 3:241–247

Received December 17, 1987/Revised and accepted March 28, 1988

Appendix

Comparison of Six DNA Sequences. Table 6 summarizes the parsimony analysis of three lysozyme and three lactalbumin sequences considered at once. The nine trees that unite the three lactalbumins with one another are greatly superior to the other 96 possible bifurcating trees (see legend to Table 6).

The number of informative sites that distinguish between two trees when the 36 possible pairs among the nine trees in the table are considered ranges from 13 (A_2 vs A_3, B_2 vs B_3, and C_2 vs C_3) to 60 (A_1 vs C_2). None of the B or C trees is statistically superior to any of the A trees, an inference that can be drawn also by noting that in the table itself the A trees all require fewer events than the other six trees. Keeping the intralactalbumin arrangement constant, the score of sites favoring A vs those favoring B or C and the associated probabilities of statistical equivalence for the six comparisons are as follows: A_1 vs B_1, 28:16, 0.048*; A_1 vs C_1, 28:15, 0.033*; A_2 vs B_2, 27:16, 0.063; A_2 vs C_2, 27:13, 0.019*; A_3 vs B_3, 29:18, 0.072; A_3 vs C_3, 29:15, 0.024*. Thus, the A trees are superior to the B and C trees at a level of statistical significance between 2% and 8%.

Table 6. Parsimony analysis of DNA sequences for animal lysozymes *c* and α-lactalbumins

Lactalbumin arrangement		Events per tree		
Pair joined	Outside species	A	B	C
1) Human + guinea pig	Rat	371	385	386
2) Human + rat	Guinea pig	374	387	390
3) Guinea pig + rat	Human	379	392	395

The numbers of mutations required to account for the observed sequence diversity at the 191 phylogenetically informative sites among the six sequences in Fig. 3 are tabulated here for the nine best (most parsimonious) trees. The remaining 96 bifurcating trees, in which the three lactalbumins do not form a cluster, require 418–503 events at the 191 sites. Deletions or additions within one codon were counted as deletions or additions of individual nucleotides for computation of the total number of mutations required by each of the 105 trees at the 191 sites; eight codons (nos. 14, *24*, 47, 100, *119*, 125, 126, and *between 127 and 128*) with gaps inferred for one or more sequences had two or three informative sites (Fig. 3), but only the three italicized codons were relevant to comparisons of the nine most parsimonious trees with one another. As in Table 3, for pairwise comparisons of trees and statistical testing, two deletions within a codon were treated as one informative site. The nine trees in the table may be referred to as A_1–A_3, B_1–B_3, and C_1–C_3, with branching orders A, B, and C corresponding to those shown in Fig. 2 and branching orders 1, 2, and 3 referring to the arrangements among the three lactalbumins indicated here

Although A_1 and A_2 are more parsimonious than A_3, which has the intralactalbumin branching order expected from organismal relationships, the differences among the three trees are not statistically significant: A_1 vs A_2, score of 11:9, $P = 0.412$; A_1 vs A_3, 11:4, 0.059; A_2 vs A_3, 9:4, 0.133. This result is consistent with our view (Sarich 1985) that the time of divergence between the rat and guinea pig lineages is nearly as ancient as that between rodents and primates.

Comparisons Using Goat α-lactalbumin. After this paper was accepted for publication, we became aware of the DNA sequence encoding goat pre-α-lactalbumin (Kumagai et al. 1987). The mature protein is 123 amino acids long and the signal peptide 19; the sequence is exactly alignable with those of human and guinea pig. In standard parsimony tests like those in Table 3, the number of sites favoring trees A, B, and C, *P* for A vs B, and *P* for A vs C are, respectively, as follows (with v standing for transversions and i for transitions and asterisks marking *P* values <0.05):

All positions:	27, 17, 13, 0.087, 0.019*
All positions except gaps:	25, 17, 13, 0.140, 0.036*
All transversions:	15, 8, 5, 0.105, 0.021*
All transitions:	10, 9, 8, 0.500, 0.407
First + second positions:	16, 7, 3, 0.047*, 0.0022*
First + second positions, v:	8, 3, 1, 0.113, 0.020*
First + second positions, i:	8, 4, 2, 0.194, 0.055
Third positions:	9, 10, 10, 0.500, 0.500
Third positions, v:	7, 5, 4, 0.387, 0.274
Third positions, i:	2, 5, 6, 0.227, 0.145

Tree A is supported just as when human, guinea pig, and rat lactalbumins are used in the comparisons.

In evolutionary-parsimony tests like those in Table 5, the scores of parsimony to background sites for the three trees are as follows:

All positions:	A, 16:13; B, 9:5; C, 7:4
First + second positions:	A, 8:5; B, 4:1; C, 2:3
Third positions:	A, 8:8; B, 5:4; C, 5:1

None of the associated *P* values is below 0.10. These results reinforce the notion that the evolutionary-parsimony method is of limited value in choosing among alternative trees in the present case.